Maternal weight trajectories around pregnancy and obesity in early childhood

Study Protocol
Version 1.1
2020-03-04

1.0. Background

Childhood obesity is becoming a more severe problem in the United States, particularly in the southeast region of the nation. Maternal behaviors and characteristics are critical for determining the health of their children during the first few years of life. There is keen interest in better understanding factors associated with childhood obesity and potential interventions to lower its risk. In particular, there is interest in understanding the relative contribution of a mother's weight before, during, and after pregnancy on childhood obesity.

The Vanderbilt University Medical Center (VUMC) includes electronic medical records data for over 2 million patients. Linked data from mothers and children delivered at VUMC are available and provide us with the opportunity to study factors associated with childhood obesity from a large and diverse group of mothers and children. There are also some challenges with using electronic health records data for medical research, because the data are often more prone to errors. We will validate targeted subsamples of electronic health records and apply novel statistical methods to improve estimation.

2.0. Study Objective:

This study will examine the relative contributions of maternal weight trajectory before, during, and after pregnancy on the cumulative probability of childhood obesity at ages 2 and 5 years. We will consider whether patterns in mother weight, such as being overweight or obese before the pregnancy, gaining excess weight during pregnancy and not losing it, gaining excess weight during pregnancy and losing it, or not gaining excess weight affect the risk of childhood obesity. Other childhood outcomes in relation to obesity of the child and mother will also be considered, such as asthma and cardiovascular outcomes (e.g. high blood pressure).

2.1. Inclusion Criteria:

Mothers and child dyads at VUMC will be included. VUMC has the ability to link mother-child EHR, and in an earlier dataset, there were 27,390 mother-child pairs with an existing linkage recorded in a Mom-Baby Linkage table.

Maternal inclusion criteria:

- In a mother-child dyad in the Mom-Baby Linkage table
- Age $\geq$ 16 years
- At least one measure of weight within 273 days prior to birth of the baby

- At least one height measure

Child inclusion criteria:
- In a mother-child dyad
- At least 2 weights, where one weight occurs within 24 hours of birth
- At least one simultaneous height and weight measure at >730 days after birth and before 6 years of age

3.0. Outcomes

Primary outcome:
Time from birth to body mass index (BMI) at or above the 95th percentile based on sex-specific reference populations using CDC BMI-for-age growth charts for children ≤5 years

Secondary Outcomes:
- Child Weight at 2 years and 5 years
- Incidence of childhood obesity at 2 and 5 years
- Asthma at age 4-5 years

Exploratory outcomes:
Childhood comorbidities, including high cholesterol, and high blood pressure, in first five years of age

4.0. Study Design
This is a retrospective, observational study using already collected data. All eligible mother-child dyads in the VUMC Mom-Baby Linkage table will be included for study. We will be prospectively validating EHR data for a subsample of patient records from this cohort, further details are provided below.

5.0. Statistical Considerations

5.1. Analysis Considerations

Key covariates: Maternal race/ethnicity, maternal age at pregnancy event, maternal diabetes, tobacco use, medications, mode of delivery, maternal mental health status, socio-economic status, parity, marital status, alcohol use, child sex, child clinical diagnoses, and estimated gestational age of the child at delivery.

Proposed analyses: The cumulative probability of childhood obesity according to age will be estimated using Kaplan-Meier methods. Follow-up for primary time-to-event analyses will be the time from turning 2 years old until the first of childhood obesity, last measurement in the EHR, or 6 years. The association between time-to-obesity and maternal weight trajectories

before, during, and after pregnancy will be examined using Cox regression adjusted for the key covariates listed above. Maternal weight trajectories will be summarized using weight during the year prior to pregnancy, weight change during pregnancy, and weight change during the year after pregnancy. Weight change will be treated as a continuous variable for all analyses, although we may consider additional analyses that categorize it.

We anticipate substantial variation in the frequency of measurements across mothers and children, and we anticipate the appropriate choice of analysis method will depend on the level of completeness of the data. We may consider joint models of longitudinal maternal weight measurements and times-to-childhood obesity. We will also consider functional analysis methods to model overall trends in addition to individual weight trajectories. We will consider very carefully methods to adjust for potential biases that could arise if the numbers of measurements on a woman or child are informative of health status. We will consider propensity score type adjustments, to adjust for the propensity of having data at a specific milestone (e.g. child 2yr or 5yr birthday), as well as methods such as outputation (Follmann et al 2003) which will equalize the number of measures per individual to explore robustness of results across different methods to adjust for these potential biases. Standard methods, e.g. multiple imputation, will be used to handle missing data.

We will assess the association between maternal weight trajectories and childhood asthma at age 4 or 5 using logistic regression, limited to the children with data at ages 4 or 5 years. Multivariable logistic regression models will adjust for similar covariates to those described above for the obesity analyses.

Methods to address error in variables: We anticipate there being a moderate number of errors in weight measurements, primarily due to measurement error, but some also due to recording errors. We anticipate that date of birth will be largely accurate in the EHR, but that other dates and clinical events will be prone to errors. Some of the covariates (specifically, race, diabetes, tobacco use, and medications) are known to be particularly error-prone, and we anticipate that patient records with errors in one variable will tend to be more likely to have errors in multiple variables. We also anticipate a moderate number of errors in the linkage of maternal-child records.

We will implement a multi-wave, targeted validation sub-study; the validation strategy is outlined in the next section. Once we have obtained validation data for a subsample of patient records, we will perform the proposed study analyses highlighted above but augmented using specific methods to incorporate both the validated and unvalidated data. We will consider multiple analysis approaches as what is optimal for the problem may depend on the error structure. Namely, we will consider a maximum likelihood, multiple imputation, and generalized raking approaches. The optimal raking variable will be constructed using multiple imputation, as described in Han et al. (2019). Single imputation methods, such as regression calibration, in combination with raking will also be considered (Oh et al., 2019) These methods will take into account the targeted nature of the validation sample (i.e., the fact that the

validation sample is not a completely random sample, but rather a targeted random sample [e.g., weighted or stratified sampling] in the first and second validation waves).

Data Validation Strategy

Our goal will be to optimize the information learned (i.e., minimize variance of corrected estimates) based on our budgetary constraints; we anticipate we will be able to validate between 500-1000 records.

Data validators will receive validation training from Drs. Shepherd, Duda, and Heerman, and will enter validation data into standardized REDCap validation forms built off of the PCORnet Common Data Model. We anticipate validation that will happen in 2 waves. The first wave will be treated as a pilot sample and information learned regarding data discrepancies will be used to tailor the individuals selected in the second validation wave.

We will design our multi-wave validation sampling with two separate analyses in mind: (1) association between maternal weight change during pregnancy and time to childhood obesity and (2) association between maternal weight change during pregnancy and probability of asthma for child. We will design our validation sample such that we develop two separate sampling schemes to optimize these two separate analyses. As childhood obesity is our primary endpoint, approximately 2/3 of our chart reviews will be targeted towards this endpoint, the other 1/3 will be targeted towards childhood asthma. Naturally, the charts that are targeted for validation to improve estimation of the childhood obesity outcome will still be used for the childhood asthma analyses and vice versa, even though they are not targeted for that purpose.

For the first wave of validation sampling, no more than half of the records will be sampled according a balanced stratified sample, with randomly selected samples within strata. For the childhood obesity endpoint, we will create 6 strata based on all combinations of obesity (yes/no) and rate of maternal weight gain ($\leq 25^{th}$ percentile, 25^{th}-75^{th} percentile, $>75^{th}$ percentile). In wave 1 of chart validation, we will validate 40 charts per stratum (total of 240 charts). For the asthma endpoint, we will create 6 strata based on all combinations of childhood asthma (yes/no) and rate of maternal weight gain ($\leq 25^{th}$ percentile, 25^{th}-75^{th} percentile, $>75^{th}$ percentile); note that the only patients included in these analyses will be those children followed with at least one visit from ages 4-5 years. In wave 1, we will validate 25 charts per stratum (total of 150). Therefore, our first wave of chart reviews will include a total of 390 charts. For the chosen records, validation will be performed by going through the entire EHR (free text and other fields not part of the PCORnet Common Data Model) to validate key study variables, which include the primary outcomes, exposures, and key covariates given above. De-identified validation data will be securely transferred to study investigators (Drs. Shepherd and Shaw) who will then use this information to tailor a second wave of data validation.

For the second wave of validation, we anticipate over-sampling/targeting records that are more influential, such as those that appear to be more prone to errors (based on our initial wave of data validation), records that are more likely to have observed events, and records with extreme values or influence on the estimating equations. We plan to target our samples to

minimize the variance of the estimate of our primary outcome (time-to-obesity) and secondary outcome (incidence of asthma). Hence, we plan to sample following adaptations of the mean-score sampling approach of McIsaac and Cook (2015) to optimize sampling for the discrete proportional hazards and logistic regression models. Neyman allocation strategies will be employed to assure efficiency and to guarantee sufficient information per stratum so that all analysis methods may be employed. If the first wave reveals a large amount of outcome misclassification, then these methods may be adjusted so as stratify on predicted true positive versus predicted true negative results, versus the error prone outcome alone in the second wave. For example, rather than sampling based on influence functions plugging in the phase 1 data, we may sample based on influence functions plugging in multiply imputation estimators based on the wave 1 validation. We anticipate validating several hundred additional records for this wave 2 validation; hopefully at least as many as in the wave 1 validation. Exact numbers to validate will be determined based on feasibility and costs learned from the wave 1 validation. Data from this second wave of validation will be similarly entered into REDCap and transferred to study investigators.

References

Follmann D, Proschan M, Leifer E. Multiple outputation: inference for complex clustered data by averaging analyses from independent data. Biometrics. 2003 Jun;59(2):420-9.

Han K, Shaw PA, Lumley T. Combining multiple imputation with raking of weights in the setting of nearly-true models. (arXiv:1910.01162[stat.ME])

McIsaac MA, Cook RJ. Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis. Statistics in Medicine. 2015 Sep 20;34(21):2899-912.

Oh E, Shepherd BE, Lumley T, Shaw PA. Raking and Regression Calibration: Methods to Address Bias from Correlated Covariate and Time-to-Event Error. (arXiv:1905.08330v1 [stat.ME])